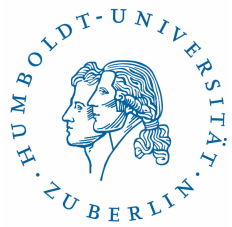
Humboldt-Universität zu Berlin
School of Business and Economics
Ledislaus von Bortkiwiecz Chair of Statistics

Regressionsanalyse des Berliner Immobilienmarktes

Regression Analysis of Berlin Housing Market



Bachelorarbeit
von
Karolina Stańczak

zur Erlangung des akademischen Grades
Bachelor of Science (B.Sc.)
in Volkswirtschaftslehre

Matrikel-Nummer 526005

Prüfer: **Prof. Dr. Wolfgang Härdle**
Betreuer: **Dr. Sigbert Klinke**

Berlin, den 19. Juli 2014

Abstrakt

In dieser Arbeit wurde ein hedonisches Preisberechnungsmodell für den Berliner Immobilienmarkt auf Basis der Immobilienverkäufe von 2003 bis 2013 erstellt. Dazu wurden aggregierte Daten für Berliner Ortsteile verwendet. Es wurden deskriptve Statistiken durchgeführt und graphisch mit Karten veranschaulicht. Es wurde gezeigt, dass das durchschnittliche Stockwerk, in dem die Wohnungen durchschnittlich liegen, Raumanzahl, Wohnfläche und Anteil der Sozialmietwohnungen an allen Wohnungen einen stark signifikanten Einfluss auf den Immobilienpreis haben, welcher hier durch den Kaufpreismedian je m^2 repräsentiert wird. Um die räumliche Autokorrelation ins Modell einzubeziehen, wurde ein räumliches Modell, das simultane autoregressive Modell, angewandt. Eine leichte räumliche Autokorrelation wurde mithilfe Moran's I bestätigt. Die beiden Modelle haben nahezu gleiche Ergebnisse geliefert.

Schlagwörter: Hedonische Methode, räumliche Regression, Berliner Immobilienmarkt

Danksagung

An dieser Stelle möchte ich mich ganz herzlich bei allen bedanken, die mich bei dieser Arbeit unterstützt und motiviert haben.

Ein besonderer Dank gilt meinem Betreuer Herrn Dr. Sigbert Klinke, der mich immer mit vielen Ideen unterstützt hat. Seine wissenschaftliche Begeisterung und Erfahrung haben mich in den letzten Monaten stets durch mein Studium begleitet. Ohne seine Hilfe wäre diese Arbeit nicht zustande gekommen.

Es gibt keine Worte, die meine Dankbarkeit für meine Familie - meine Eltern und meinen Bruder - beschreiben könnten. Sie haben mir nicht nur das Studium ermöglicht, sondern mich auch bei jeder Entscheidung unterstützt haben. Zuletzt ein großes Dankeschön an alle meine Freunde, ohne sie wäre das Studium nur halb so fantastisch gewesen. Ihr wisst, wie dankbar ich bin.

Inhaltsverzeichnis

Abbildungsverzeichnis

Tabellenverzeichnis

1 Einleitung

Immer öfter wird Berlin in den Medien, nicht nur in Deutschland, sondern auch weltweit, als die neue internationale Hauptstadt Europas genannt. Berlin ist zur Zeit das Bundesland mit dem größten wirtschaftlichen Wachstum. Internationale Großunternehmen und einheimische Start-Ups eröffnen Zentralen und Filialen in Berlin. Somit entstehen immer mehr neue Arbeitsplätze und die Arbeitslosenquote sinkt. Dies verursacht sehr hohe Zuwanderungsgewinne (über 40.000 im Jahr 2012). Diese Situation belastet den Berliner Immobilienmarkt. In den letzten Jahren erfüllt die Bautätigkeit die Bedürfnisse des Marktes nicht, wodurch sich das Wohnungsdefizit vergrößert. Die steigenden Mietpreise in Berlin sind eines der wichtigsten Themen in der politischen Debatte der Stadt. Obwohl Berlin mit einem Anteil selbst genutzten Wohneigentums von lediglich 15%[1] eine Mieterstadt ist, beeinflussen sich die Miet- und Kaufpreise gegenseitig und betreffen so alle Bürger. Die Angebots-Kaufpreise für Eigenwohnungen sind im Jahr 2012 um 14% gestiegen.[2]

Der Immobilienmarkt ist ein charakteristisches Beispiel für einen heterogenen Markt, auf dem jedes Gut - auch eine Immobilie - einzigartig bezüglich ihrer Charakteristiken ist. Dies führt zu einer assymetrischen Information zwischen den Marktbeteiligten und erhöht die Suchkosten. Außerdem kann sich der Immobilienmarkt erst mit großen Verzögerungen an die allgemeine Wirtschaftssituation anpassen, da die Anfertigung einer Immobilie eine gewisse Zeit benötigt. Deswegen ist das Immobilienangebot kurzfristig gesehen sehr unelastisch, jedoch ist die langfristige Preiselastizität der Immobilien relativ hoch. Wegen zuvor entstandener Überkapazititäten auf dem Berliner Immobilienmarkt, wurde im letzten Jahrzehnt das Naubauvolumen auf einem konstant niedrigen Niveau gehalten. Im Jahre 2012 hat sich diese Situation geändert und die Anzahl der Baugenehmigungen ist im Vergleich zum Vorjahr um 35% gestiegen, um die schon seit Jahren steigende Wohnungsnachfrage zu decken.

Um ein besseres Verständnis vom Berliner Immobilienmarkt zu bekommen, werden in dieser Arbeit die Kaufpreise der Eigentumswohnungen analysiert. Es wird versucht, den Preis in seine Bestandteile zu zerlegen. Dieser Ansatz liegt der hedonischen Methode zugrunde, die im Rahmen der Arbeit erläutert wird. Die Charakteristiken der Berliner Immobilien werden beschrieben und graphisch dargestellt, genauso wie die Charakteristiken der Ortsteile. Diese Charakteristiken sollen schließlich als Kovariaten eines Regressionsmodells die Immobilienpreise in den Berliner Ortsteilen erklären.

Der Datensatz, der für die Zwecke dieser Arbeit erstellt wurde, basiert auf dem Boston Housing-Datensatz, der die Wohnverhältnisse in 506 Einschulungsgebieten von Boston auf Basis einer Erhebung aus dem Jahr 1970 beschreibt. Für jedes der Gebiete wurden Werte für 13 Variablen ermittelt und der Hauspreismedian wurde in der ursprünglichen Analyse als die abhängige Variable untersucht. Dieser Arbeit liegt die Idee einer solchen Datenbank für Berlin zugrunde. Diese Datenbank sollte unterschiedliche Variablen beinhalten, wie Wohnungsattribute, Nachbarschaftsatributte und sozio-ökonomische Charakteristiken. Der Boston Housing-Datensatz und diesem ähnliche haben unter Forschern ein großes Interesse geweckt. Insbesondere ist hier die Arbeit von D. Harrison und D. Rubinfeld "Hedonic Housing Prices and the Demand for Clean Air" gemeint, die meine

[1] nach Zensus 2011

[2] Berliner Wohnungsmarktbericht 2013

Aufmerksamkeit auf das Thema der hedonischen Immobilienpreisanalyse gerichtet hat. Die Idee der Untersuchung der räumlichen Beziehungen zwischen den Berliner Ortsteilen wurde dank der Arbeit von W. Brunauer, S. Lang, P. Wechselberger und S. Bienert Ädditive Hedonic Regression Models with Spatial Scaling Factors: An Application for Rents in Vienna" und deren räumliche Analyse der Mietpreise in Wien ausgelöst.

In dieser Arbeit werden zuerst der Datensatz und die Schritte, die bei der Erstellung benötigt wurden, beschrieben. Danach werden alle Methoden, die bei der Anfertigung dieser Arbeit verwendet worden, erklärt. Als nächstes werden die deskriptiven Statistiken zu den in dem Datensatz erhaltenen Variablen untersucht. In Kapitel 5 werden Regressionsanalysen durchgeführt. Als Erstes wird der Ansatz des additiven Regressionsmodells angewendet und im nächsten Schritt liegt der Fokus auf den räumlichen Beziehungen, welche in der räumlichen Regression betrachtet werden. Schließlich werden die Annahmen überprüft und eine Entscheidung bezüglich der Modellwahl wird getroffen, um die Schlussfolgerungen vorstellen zu können.

2 Datensatz

Diese Arbeit stellt den Versuch dar, den Preis einer Immobilie in einem Ortsteil in seine Einzelteile zu zerlegen. Erstens sind das die Eigenschaften der Wohnung selbst, sowie die Wohnfläche und das Stockwerk, und zweitens die Eigenschaften der Nachbarschaft - Verkehrsanbindung, Kriminalität, oder die Wohnlage. Die letzte Komponente stellen die Umweltcharakteristiken dar, zu denen beispielsweise Parkflächen in der Umgebung oder Luftverschmutzung zählen. Diese Eigenschaften werden in diesem Projekt für die Berliner Ortsteile zusammengefasst, um weiterhin mit einem Datensatz zu arbeiten, der die Wohnverhältnisse in Berlin beschreibt. Dies ist ein ähnlicher Ansatz zu dem bekannten, in 70er Jahr enstandenen, Boston Housing-Datensatz.[3] Zunächst wird die kleinräumige Gliederung Berlins nahegebracht und die Schritte der Datensammlung und -aufbereitung werden erläutert.

2.1 Kleinräumige Gliederung

Abbildung 1: Berliner Ortsteile

Innerhalb dieser Arbeit wird mit kleinräumigen Daten gearbeitet und somit kann man sich die kleinräumige Gliederung Berlins zu Nutze machen. Berlin wird in zwölf Bezirke unter-

[3] Vgl. [6]

teilt. Eine weitere Unterteilung, die lebensweltlich orientierten Räume (LOR), wurde von der planenden Fachverwaltung des Senats und dem Amt für Statistik Berlin-Brandenburg erfasst. Die LOR bestehen aus 3 Ebenen: 447 Planungsräume, 138 Bezirksregionen und 60 Prognoseräume. Eine andere Möglichkeit der Unterteilung geben die 95 (seit 2012, nach der Abgrenzung Borsigwalde von Wittenau - 96) Ortsteile.[4] Eine Liste aller Berliner Ortsteile mit den zugehörigen Ortsteilnummern befindet sich im Anhang in der Tabelle 10, was in der Abbildung 1 grafisch dargestellt wurde.

2.2 Immobiliencharakteristiken

Die Daten zur Immobiliencharakteristik wurden von dem Gutachterausschuss für Grundstückswerte in Berlin (GGA) für Zwecke der Arbeit zur Verfügung gestellt. Der GGA ist gesetzlich durch den § 195 des Baugesetzbuches (BauGB) dazu verpflichtet, eine Kaufpreissammlung jedes Vortrags zur Übereignung von Eigentum an einer Immobilie gegen Entgelt zu führen[5]. Die Daten zur Kaufpreissammlung werden um die Informationen zur Grundstücksbewirtschaftung und anderen preis- und wertrelevanten Faktoren ergänzt. Diese Daten werden in die Automatisierte Kaufpreissammlung Berlin (AKS Berlin) übernommen. Es ist möglich die Informationen über den Berliner Grundstücksmarkt als Grundlage für Marktanalysen herunterzuladen, oder sie zu bestellen, was für wissenschaftliche Untersuchungen unentgeltlich ist.

In dem bestellten Datensatz befinden sich für den Zeitraum von Januar 2003 bis Dezember 2013 Angaben zu allen zwölf Berliner Bezirken. Die 95 Variablen nehmen sowohl die quantitativen - das Baujahr, den Kaufpreis je m^2, die Wohn- und Nutzfläche und die Anzahl der Räume - als auch die qualitativen Merkmalsausprägungen an, sowie stadträumliche Wohnlage, die Wohnungsart und den Orsteil, in dem sich die Immobilie befindet. Die Variable stadträumliche Wohnlage aus dem Datensatz von GAA nimmt folgende Merkmalsausprägungen an: einfache, mittlere, gute und sehr gute Wohnlage - je nach Babauung, Straßenbild, Durchgrünung und Verkehrsanbindung. Als sehr gute Wohnlage wurden nur wenige Gebiete (ausschließlich im Westteil Berlins) definiert, wie Umgebung der Heerstraße, Schmargendorf und Grunewald.

2.3 Nachbarschaftcharakteristiken

Die Bestandteile der Nachbarschaftcharakteristiken, mit denen es gearbeitet wird, sind Daten zur stadträumlichen Wohnlage, der Flächennutzung, der Anzahl der Haltestellen, der Straftaten, der Anzahl der Schulen, zu dem Anteil der Sozialwohnungen und Daten aus dem Sozialstrukturatlas - Anteil der Population unter 18 Jahre, Anteil der Ausländer und der Arbeitslosen.

Die Informationen zur Flächennutzung kommen aus dem FIS-Broker, einer von der Senatsverwaltung für Stadtentwicklung und Umwelt zur Nutzung der Berliner Geodaten entwickelten Online-Plattform. In dem Datensatz befinden sich 26 Variablen. Diese stellen dar, wie die Nutzung des jeweiligen Prognoseraums aufgeteilt ist, insbesondere die Straßen, sowie Park- und Grünflächen in m^2. Vom FIS-Broker kommen ebenso Daten

[4] Da es meist mit Daten vor 2012 gearbeitet wird, wird mit der Unterteilung vor 2012 gearbeitet.

[5] siehe `https://www.gutachterausschuss-berlin.de/gaaonline/index.html`

zum Thema Ausbildung, denn sie beinhalten alle Schulen in Berlin mit Berücksichtigung des Ortsteils. Da die Verteilung und Menge der Schulen einen Bedeutenden Einfluss auf die Attraktivität der Umgebung für junge Familien haben könnte, wird auch diese Annahme in der Arbeit geprüft. Der FIS-Broker liefert außerdem Informationen darüber, welchen prozentualen Anteil an allen Wohnungen Sozialwohnungen in jeweiligen Prognoseräumen stellen.

Das Center Nahverkehr Berlin (CNB) hat die Daten zur Anzahl der Haltestellen geliefert. CNB ist ein Team, das im Auftrag der Senatsverwaltung für Stadtentwicklung arbeitet und sich mit der Kontrolle und Optimierung des Nahverkehrs beschäftigt. Die verwendeten Daten geben Auskunft über die Anzahl aller Haltestellen in Berlin (Tram, Bus, S-bahn und U-bahn) und der Anzahl der S-bahn- bzw. U-bahnstationen je Prognoseraum.

Die Daten zu Straftaten in den Berliner Ortsteilen kommen von der Polizei aus der Polizeilichen Kriminalistatistik (PKS), einer statistischen Zusammenstellung der der Polizei bekannt gewordenen strafrechtlichen Sachverhalte. Eine von der PKS erstellte Statistik ist der Kriminalitätsatlas, aus dem die benötigten Daten abgeleitet wurden (Ausgabe 2011). Es wurden alle erfassten Straftaten bekannt gegeben, die häufigsten Straftaten wurden auch separat angegeben (darunter Raub, Einbruch, Diebstahl und ähnliches). Für die Zwecke dieser Arbeits wird mit den Häufigkeitszahlen (Zahl der bekannt gewordenen Fälle errechnet auf 100.000 Einwohner) zu Straftaten im Allgemeinen gearbeitet.

Aus dem Sozialstrukturatlas 2013 (Stichtag: 31.12.2011) kamen die Daten bzgl. des Anteils der Personen unter 18 und des Anteils der Ausländer und Arbeitslosen in den jeweiligen Prognoseräumen, die in Prozentzahlen angegeben wurden. Da diese soziodemographischen Merkmale angemessene Attribute zur Nachbarschaftcharakteristik zu sein scheinen, werden sie in dem Datensatz ebenso berücksichtigt.

2.4 Datenaufbereitung

Die gesammelten Daten wurden wie folgt aufbereitet, um letztendlich einen umfangreichen Datensatz bezüglich Wohnverhältnisse in den Berliner Ortsteilen zu bilden.

Da der Datensatz zu den Berliner Schulen lediglich eine Liste ist, wurden die Lehranstalten in den jeweiligen Ortsteilen zusammenaddiert, um die Anzahl der Schulen je Ortsteil zu bekommen, mit der zunächst gearbeitet wird. Die Daten zur Flächennutzung beziehen sich auf die Ebene der Planungsräume, deswegen wurden sie aggregiert, um die Daten zu den Prognoseräumen zu bekommen.

Ein nächster essenzieller Punkt in der Vorbereitungsphase war die Vereinheitlichung der Bezugsebene für die verwendeten Variablen. Ein Teil der Daten, nämlich die Flächennutzung, die Azahl der Haltestellen, der Anteil der Sozialwohnungen, der Personen unter 18, der Ausländer und der Arbeitslosen, bezog sich auf die Planungs- bzw. Prognoseräume. Die restlichen Daten bezogen sich auf die Berliner Ortsteile. Wegen dieser Uneinheitlichkeit mussten die Daten zunächst auf die selbe Bezugsebene - die Ortsteile - gebracht werden. Dies kam zu stande mit Hilfe einer gewichteten Umrechnung durch die Wohnungszahlen der Ortsteile und für die Teilstücke der Ortsteile, die durch Prognoseräumen entstehen, zustande. Die benötigten Tabellen wurden von dem Amt für Statistik Berlin-Brandenburg geliefert. Die erste der erhaltenen Tabellen beinhaltet Informationen zum

Anteil der Wohnungen eines Ortsteils am Prognoseraum[6] und war bei der Umrechnung der Anzahl der Haltestellen und der Flächennutzung hilfreich. Die zweite Tabelle (zum Anteil der Wohnungen eines Prognoseraums am Ortsteil) hat die Umrechnung des Anteils der Sozialwohnungen, der Personen unter 18, der Ausländer und der Arbeitslosen ermöglicht.

Im nächsten Schritt wurden aus den Daten aus dem Immobiliendatensatz Variablen zum Kaufpreismedian je m^2, dem Anteil der Altbauten (bis 1949) an verkauften Wohnungen, den durchschnittlichen Werten für ein Stockwerk (in dem die Wohnungen in den jeweiligen Ortsteilen durchschnittlich liegen), die Anzahl der Räume, die stadträumliche Wohnlage und die Wohnfläche für die jeweiligen Ortsteile erstellt. Die Daten zur Flächennutzung, namentlich Straßen und Parkfläche in m^2, wurden durch die Gesamtfläche der Ortsteile dividiert, um die prozentualen Anteile der Straßen bzw. Park- und Grünflächen zu ermitteln. Danach wurden alle Variablen zu den Berliner Ortsteilen bzgl. der Immobilien- und Nachbarschaftcharakteristik zu einer Datenbank, welche eine Basis für weitere Analysen der Wohnverhältnisse in Berlin ist, zusammengeführt. Alle Variablen sind mit Erklärung in der Tabelle1 zu finden.

Variable	Erklärung
KP	Kaufpreismedian je m^2
Baujahr	Anteil der Altbauten (bis 1949) an allen Gebäuden
Ge	Durchnittliches Stockwerk der Wohnung
Ra	Durchschnittliche Raumanzahl
SW	Durchschnittliche stadträumliche Wohnlage
WF	Durchschnittliche Größe der Wohnung in m^2
Strassen	Anteil der Strassen an der Gesamtfläche
Park	Anteil der Park- und Grünfläche an der Gesamtfläche
Haltestellen	Anzahl der Bus, Tram, Ubahn und Sbahn Haltestellen
Ubahn	Anzahl der Ubahn-Stationen
Sbahn	Anzahl der Sbahn-Stationen
Straftaten	Häufigkeitszahl des Straftaten insgesamt
Schulen	Anzahl der Schulen jeder Art
Sozialmietwohnungen	Anteil der Sozialmietwohnungen an allen Wohnungen
unter18	Anteil der Population unter 18
ausland	Anteil der Ausländer
arbeitslos	Anteil der Arbeitslosen

Tabelle 1: Liste der Variablen.

Das räumliche Modell und alle graphischen Darstellungen auf den Karten basieren auf der Datei im kml-Format, die von der Senatsverwaltung für Stadtentwicklung und Umwelt veröffentlicht wurde.[7] Dateien in der Keyhole Markup Language (KML) beschreiben Daten für Google Earth und Google Maps und ermöglichen die Modellierung von Punkten, Linien bzw. Polygonen. Die vorliegende Datei beinhltet die kleinräumige Ebene

[6] Stichtag: am 09. Mai 2011; Bei Gebäuden wurden berücksichtigt: Wohngebäude, sonstige Gebäude mit Wohnraum, Wohnheime; Nicht enthalten sind gewerblich genutzte Wohnungen, Diplomaten-/Streitkräftewohnungen.

[7] `http://daten.berlin.de/datensaetze/geometrien-der-ortsteile-von-berlin-stand-072012`; letzter Besuch: 18.07.2014

der Ortsteile. Um mit diesen Geometrien in R arbeiten zu können, wurde die Datei ins Shapefile-Format übertragen, was mit Hilfe eines Online-Konverters durchgeführt wurde. In der Shapefile-Datei wurde Buckow in zwei Teile unterteilt, Buckow I (der westliche Teil) und Buckow II (der nordöstliche Teil). Deswegen wurde Buckow in dem erstellten Datensatz zweimal erfasst, jedoch nur einmal in der Regression berücksichtigt.

3 Methoden

Diese Arbeit ist ein Versuch, die Einflussfaktoren auf die Immobilienpreise, insbesondere den Kaufpreismedian je m^2 in den jeweiligen Ortsteilen, zu untersuchen. Diese Analyse benötigt eine Reihe von statistischen Werkzeugen, die essenziell sind, um zunächst eine additive Regression und folgend die räumliche Regression anzuwenden. Die beiden Methoden werden den Einfluss der Variablen auf die Zielgröße überprüfen. Zuerst wird jedoch die hedonische Mehtode ausgeführt, die der vorherigen Unterteilung der Variablen auf Immobilien- und Nachbarschaftcharakteristiken zugrunde liegt.

3.1 Hedonische Methode

Die Theorie, die der hedonischen Regression zugrunde liegt, wird Rosen (1974) zugeschrieben. Die hedonische Methode besagt, dass der Preis einer Immobilie durch jeweilige Zusammensetzung der Bündel ihrer Charakteristiken beeinflusst wird. Anders ausgedrückt ist der Preis einer Immobilie eine Summe der Preise für ihre Charakteristiken und jede Charakteristik hat einen eigenen impliziten Preis (vgl. Brasington (2005)). Dies führt zum nächsten Gedanken, nämlich dass die Transaktion auf dem Immobilienmarkt eigentlich ein Bündel von Transaktionen ist. Für das hedonische Modell wurde keine eindeutige Formel definiert und deren Bestimmung hängt vom Anwender ab. Eine der Möglichkeiten der Gleichung kann wie folgt formell beschrieben werden:

$$E(Y|X_1 = x_1, \ldots, X_D = x_D) = \alpha + \sum_{d=1}^{D} m_d(x_d).^{8} \tag{3.1}$$

Die hedonischen Preisfunktionen werden also wie multivariaten Regressionsmodelle modelliert. Die Werte x_i messen die Preise der Charakteristiken und die Werte $m_d(.)$ der Funktionen auf den Werten von x_i. Aus den beobachteten Werten der Wohnungsattibute können die Preise für die jeweiligen Charakteristiken ermittelt werden. Dies wird mithilfe einer Regression erzielt. Die Koeffizienten der erklärenden Variablen zeigen die impliziten Preise für die jeweiligen Attribute. Das Preisniveau hängt von der Nachfrage und dem Angebot ab und spiegelt wider, welche die jeweiligen Charakteristiken gefragt sind.

Die hedonische Methode kann natürlich nicht nur im Fall von Immobilien angewendet werden. Dieses Verfahren wird auch bei allen anderen Gütern angewendet, deren Preis von ihren Attributen beeinflusst wird, z.B. bei Computern, Autos, Kleidung. Für Computer sind die Attribute u.a. die Geschwindigkeit des Prozessors, die Grösse der Festplatte und die des Arbeitsspeichers. Im Fall einer Immobilie sind das die Grösse der Wohnung, die Anzahl der Räume und die Entfernung zu dem nächsten Park. In dieser Arbeit werden, da mit aggregierten Daten zu den Berliner Ortsteilen gearbeitet wird, die Charakteristiken der einzelnen Wohnungen durch den Kaufpreismedian je m^2 und die durchschnittlichen Werte der Attributen in den Ortsteilen repräsentiert.

[8] Vgl. Martins-Filho C. und Bin O. (2005)

3.2 Das Regressionsmodell und die KQ-Schätzung

Das Hauptziel einer Regression ist, eine abhängige Variable durch eine oder mehrere erklärende unabhängige Variablen zu beschreiben. Die Schätzung der Regressionskoeffizienten erfolgt üblicherweise mit Hilfe der Methode der kleinsten Quatrate, die nachfolgend erläutert wird.

3.2.1 Das additive Regressionsmodell

Einer der Fälle der Regressionsmodelle und gleichzeitig eine Erweiterung des linearen Modells, welche mehrere erklärende Variablen in Betracht zieht und eine nichtlineare Beziehung einer oder mehreren Variablen mit der Zielgröße zulässt, ist das additive Regressionsmodell, das wie folgt beschrieben wird:

$$y_i = f_1(z_{i1}) + \cdots + f_q(z_{iq}) + \beta_0 + \beta_1 x_{i1} + \cdots + \beta_k x_{ik} + \varepsilon_i. \tag{3.2}$$

Der Wert der j-ten Kovariable ist gleich x_{ij} für die i-te Beobachtung, wobei die Anzahl der Bobachtungen $i = 1, \ldots, n$ ist. Dabei sind die Werte β_k, die Regressionskoeffizienten und ε_i die entsprechenden Störterme, so wie in der linearen Regression. Zusätzlich werden die stetigen Kovariaten $z_i, \ldots, z_q$ benutzt. Die Funktionen $f_1(z_{i1}), \ldots, f_q(z_{iq})$ entsprechen den nichtlinearen Effekten der metrischen Kovariablen $z_i, \ldots, z_q$.

In der empirischen Forschung ist es möglich, dass ein Wert einer Variable von der Größe einer anderen abhängt. Das Problem einer nicht linearen Beziehung zwischen der Zielgröße und den erklärenden Variablen lässt sich oft durch eine semi-logarithmische Transformation, insbesondere in das log-lineare Modell, beheben. Jenes wird beschrieben durch:

$$\ln y_i = f_1(z_{i1}) + \cdots + f_q(z_{iq}) + \beta_0 + \beta_1 x_{i1} + \cdots + \beta_k x_{ik} + \varepsilon_i. \tag{3.3}$$

Nachdem ein Modell erstellt wurde, werden die Modellannahmen überprüft, die für das additive Modell gleich denen aus der multiplen linearen Regression sind. Insbesondere ist darunter eine Residualanalyse gemeint, mit Berücksichtigung der Normalität der Fehlerterme und Testen auf Homoskedastizität. Zuerst wird jedoch auf das mögliche Vorliegen einer Mulitikollinearität geprüft.

Das Problem einer Multikollinearität trifft bei einer Abhängigkeit auf zwei oder mehr Variablen in einer Regressionsanalyse zu. In einem solchen Fall ist die Standardannahme, dass der Rang der Regressormatrix X kleiner als K (Anzahl der Regressoren) ist. Somit ist $X'X$ (nahezu) singulär und es existieren unendlich viele Schätzungen für die Regressionskoeffizienten β. Dadurch ist β nicht identifizierbar und die Schätzungen können instabil werden. Die Überprüfung dieser Annahme kommt mithilfe des Varianzinflationsfaktor zustande. Jenes wird beschrieben durch:

$$VIF_j = \frac{1}{1 - R_j^2}. \tag{3.4}$$

Falls $\max_j(VIF_j) \leq 5$ ist, kann angenommen werden, dass das Problem einer Multikollinearität nicht vorliegt.

Eine weitere Annahme einer Regressionsanalyse ist eine Normalverteilung der Residuen. Sie kann auf zwei Arten überprüft werden: erstens durch Betrachtung einer grafischen Darstellung, z.B. Q-Q-Plots und zweitens durch Durchführung geeigneter Tests, z.B. der Shapiro-Wilk-Test auf Normalverteilung. Die Nullhypothese nimmt an, dass eine Normalverteilung vorliegt. Dies wird mit der folgenden Teststatistik überprüft:

$$W = \frac{B^2}{(n-1)S^2}, \tag{3.5}$$

wobei S^2 der Varianzschätzer der Beobachtungen ist und B^2 der Varianzschätzer unter Normalität. Die Einträge in der B^2 Matrix, b^2, sind von $x_{(i)} = \mu + \sigma^2 q_i$, mit $q_i = \phi^{-1}\left(\frac{i-3/8}{n+1/4}\right)$. Eine Verletzung der Annahme der Normalverteilung hat jedoch keinen Einfluss, weder auf die Konsistenz, noch auf Effizienz der Schätzung. In großen Stichproben sollte die Verteilung gegen die Normalverteilung konvergieren. Wenn das hier der Fall ist, sind die t- und F-Tests asymptotisch gültig.

Eine Heteroskedastizität liegt vor, wenn die Residuen ungleichmäßig gestreut sind, was heißt, dass die Varianz der Residuen keine konstante Varianz aufweist. In dem Fall nimmt die graphische Darstellung der Residuen eine charakteristische Trichterform an. Um die Aussage über Homoskedastizität zu bekräftigen, führt man üblicherweise einen Breusch-Pagan-Test Test auf Homoskedastizität durch. Die Idee, die dem Breusch-Pagan-Test zugrunde liegt, sieht folgendermaßen aus:

$$\sigma^2_{U_k} = \sigma^2 f\Big(\beta_0 + \sum_j \beta_j \tilde{z}_{jk}\Big), \tag{3.6}$$

wobei die $\tilde{z}_j$ die Variablen sind, von denen die Heteroskedastizität abhängt sind. Regressiere weiterhin:

$$\hat{u}^2_k = a_0 + \sum_{j=1}^{J} a_j z_{jk}. \tag{3.7}$$

Die Teststatistik für den Breusch-Pagan-Test unter der Nullhypothese $H_0 : R^2 > 0$ lautet $nR^2 \sim \chi^2_J$.

3.2.2 Methode der Kleinsten Quadrate

Sowohl die Regressoren x_{ik}, als auch die beobachteten Werte der Zielgröße y_i sind deterministisch. Das Prinzip der Methode der kleinsten Quadrate besteht darin, dass die unbekannnten Regressionskoeffizienten β_k durch Minimierung der Summe der quadrierten Abweichungen geschätzt werden:

$$KQ(\beta) = \sum_{i=1}^{n}(y_i - x_i)^2 = \sum_{i=1}^{n}\varepsilon_i^2 = \varepsilon'\varepsilon. \tag{3.8}$$

Um das Minimum der Funktion zu finden, leitet man den Ausdruck nach β ab und setzt dies gleich 0. Danach erhält man den KQ-Schätzer:

$$\hat{\beta} = (X'X)^{-1}X'y. \tag{3.9}$$

Nachdem die Schätzungen für die Regressionskoeffizienten bestimmt wurden, wird untersucht, ob einer der Datenpunkte einen großen Einfluss auf die Regressionsnalyse hat und somit ein Ausreißer in dem Modell ist. Jenes kann anhand der Cook's Distanz überprüft werden:

$$C_i = \frac{||\hat{y} - \hat{y}_{-i}||^2}{K\hat{\sigma}} = \frac{(\hat{\beta} - \hat{\beta}_{-i})'X'X(\hat{\beta} - \hat{\beta}_{-i})}{K\hat{\sigma}} = \frac{1}{K}\tilde{e}_i^2 \frac{p_{ii}}{1 - p_{ii}}, \tag{3.10}$$

wobei $\hat{y}_{-i} = X\beta_{-i}$ den Vektor der angepassten Werte bezeichet, ohne die i-te Beobachtung zu berechnen. Es wird somit eine gewichtete Distanz zwischen den Anpassungen von y berechnet, mit und ohne Beobachtung i. Die Faustregel besagt, dass die Werte $C_i > 4/n$ einen großen Einfluss ausüben und deswegen aus dem Modell entfernt werden sollten.

3.2.3 Maximum-Likelihood-Methode

Die Maximum-Likelihood-Methode ist das nächste statistische Verfahren zur Bestimmung der Modellparameter. Das Prinzip der ML-Schätzung ist, dass die Parameter so gewählt werden, dass deren Verteilung als Realisierung der vorliegenden Daten am wahrscheinlichsten zu sein scheinen. Dabei wird die Dichtefunktion der Fehlerterme betrachtet:

$$f(\varepsilon) = \Big(\frac{1}{2\pi\sigma^2}\Big)^{\frac{n}{2}} \exp\Big(-\frac{1}{2\sigma^2}(y - X\beta)'(y - X\beta)\Big). \tag{3.11}$$

Wenn angenommen wird, dass die Fehlerterme normalverteilt sind, kann die Formel umgeformt werden. Es gilt dann:

$$f(\varepsilon_i) = \Big(\frac{1}{2\pi\sigma^2}\Big)^{\frac{1}{2}} \exp\Big(-\frac{1}{2\sigma^2}\varepsilon_i^2\Big). \tag{3.12}$$

Aus technischen Gründen wird üblicherweise nicht die Formel 3.12 maximiert, sondern ihre logarithmierte Form, die sogenannte Log-Likelihood Funktion. Da der Logarithmus eine streng monotone Funktion ist, besitzt die logarithmierte Form das Maximum in dem selben Punkt, wie die ursprüngliche Funktion.

$$\ln L = \sum_{i=1}^{n} \ln f(\varepsilon_i) = -\frac{n}{2}\ln 2\pi - \frac{n}{2}\ln\sigma^2 - \frac{1}{2\sigma^2}(y - X\beta)'(y - X\beta). \tag{3.13}$$

Nachdem die Gleichung nach β und σ abgeleitet wurde und die Ableitungen gleich Null gesetzt wurden, erhält man:

$$\hat{\beta}_{ML} = (X'X)^{-1}X'y = \hat{\beta}_{KQ}. \tag{3.14}$$

Es fällt gleich auf, dass der ML-Schätzer gleich dem KQ-Schätzer ist. Jedoch ist der ML-Schätzer der beste unverzerrte Schätzer, wobei der KQ-Schätzer nur in der Klasse der linearen Schätzer der beste ist.

3.3 Räumliches Modell

Viele Datensätze neben der Zielgröße und den Kovariablen $(y_i, x_{i1}, \ldots, x_{ik}, \ldots, z_{i1}, \ldots, z_{iq})$ beinhalten zusätzlich Informationen zu der klenräumigen Ebene, auf der die Daten gebunden sind, z.B. durch die Koordinaten, Adresse, Postleitzahl, usw. Dies ermöglicht unter anderem, die Nachbarschaften zwischen den Gebieten zu berücksichtigen und den Einfluss eines Ortes auf die Zielgröße zu bestimmen. Außerdem bietet sich dabei die Gelegenheit die Daten grafisch auf einer Karte darzustellen.

3.3.1 Nachbarschaften

Die geoadditiven Modelle basieren auf der Annahme, dass die räumlichen Koordinaten bzw. die Lokation, Einflussgrößen in einer Regression sein können, und berücksichtigen die Korrelationen zwischen den benachbarten Regionen. Die Koordinaten werden auf einer kontinuierlichen Skala gemessen, können als quasi-stetige Variablen interpretiert werden und gelten als stetige Lokationsvariablen. Dagegen können die diskreten Lokationsvariablen lediglich eine diskrete räumliche Information beinhalten, in der Form von Polygonen mit bestimmten Grenzen, wie Bundesland oder Bezirk, auf welche sich die Daten beziehen. Im Fall der vorhandenen Daten sind das die Berliner Ortsteile.

Im Fall einer stetigen räumlichen Information werden die Nachbarschaften als der Abstand zwischen zwei Lokationen, die euklidische Distanz, interpretiert. Dies ist aber für diskrete Lokationen nicht möglich und somit benötigt die Bestimmung der Nachbarschaften einige zusätzlichen Maßnahmen. Der essenzielle Schritt in einer räumlichen Analyse ist die Bestimmung der geographischen Gewichtungen. Erstens wird das Kriterium gewählt, nach dem man die Nachbarschaften ermittelt. Anders ausgedrückt wird definiert, welche Beziehungen zwischen den Beobachtungen keinen Nullwert annehmen. Zweitens werden den Verknüpfungen zwischen den Lokationen Gewichtungen zugewiesen.

Laut einem der Nachbarschaftskriterien wird eine Liste mit der Länge n mit den Indexnummern der benachbarten Polygonen der jeweiligen Komponenten erstellt. Dies bedeutet, wenn eine Beobachtung keine Nachbarn hat, enthält deren Komponente eine Null. Die räumlichen Gewichtungen hingegen können als eine Liste der nach der Liste der Nachbarn indexierten Gewichtungen interpretiert werden. Dabei ist die Gewichtung der Verknüpfung zwischen i und j die k-te Komponente der i-ten Gewichtungskomponente und k zeigt, welcher der i-ten Werte aus der Liste der Nachbarn gleich j ist. Das heißt, wenn j keine Komponente der Liste der Nachbarn ist, ist sie kein Nachbar von i. Deswegen sind die Werte w_{ij} aus der Matrix der Gewichtungen $\mathbf{W}$ gleich null, wenn j kein Nachbar von i ist. Eine andere Möglichkeit ist, eine Liste der k-nächsten Nachbarn für jede Beobachtung. Dieses Kriterium berücksichtigt die Dichte der räumlichen Einheiten. Eine derGewichtungmatrixarten entsteht, indem die Gewichtungen so standardisiert werden, dass sie sich für jede räumliche Einheit zu eins zusammenaddieren.

3.3.2 Räumliche Autokorrelation und Moran's I

Nachdem eine Gewichtungsmatrix ermittelt wurde, wird auf eine räumliche Autokorrelation geprüft. Eine räumliche Autokorrelation misst die Korrelation einer Beobachtung mit sich selbst in einem geographischen Raum. Sie misst die Wahrscheinlichkeit, zu der

ein Ereignis in einem benachbarten Raum auftreten kann. Wenn es systematische Muster in der räumlichen Verteilung gibt, liegt eine räumliche Korrelation vor. Eine Untersuchung auf eine räumliche Autokorrelation ist essenziell, da oft eine Unabhängigkeit der Beobachtungen angenommen wird. Diese Autokorrelation ist eine Verletzung einer der Standardannahmen, insbesondere in der kleinste-Quadrate-Schätzung.

Eine Möglichkeit die räumliche Autokorrelation zu testen, gibt Moran's I. Dies ist definiert als ein Verhältnis des Produktes der untersuchten Variable und deren räumlicher Verzögerung zu dem Kreuzprodukt der untersuchten Variable, angepasst an die räumlichen Gewichtungen. Jenes wird beschrieben durch:

$$I = \frac{n}{\sum_{i=1}^{n}\sum_{j=1}^{n} w_{ij}} \frac{\sum_{i=1}^{n}\sum_{j=1}^{n} w_{ij}(y_i - \overline{y})(y_j - \overline{y})}{\sum_{i=1}^{n}(y_i - \overline{y})^2}. \tag{3.15}$$

Dabei ist y_i die i-te Beobachtung, $\overline{y}$ der Durchschnitt der untersuchten Variable und w_{ij} die räumliche Gewichtung der Vernetzung zwischen i und j. Negative Werte des Moran's I implizieren eine negative Autokorrelation, die positiven eine positive Autokorrelation und 0 ein zufälliges Muster. Um eine Signifikanz dieser Autokorrelation zu testen, wird der erhaltene Wert standardisiert: $Z(I) = \frac{I - E(I)}{S_{E(I)}}$. Das übliche Signifikanzniveau von 5% angenommen, werden die Werte der Teststatistik, die betragsmäßig größer als 1,96 sind, als signifikant gelten. Um die Werte des lokalen Moran's I für die einzelnen Regionen zu ermitteln, benötigt man eine geänderte Formel:

$$I_i = \frac{(y_i - \overline{y})\sum_{i=1}^{n} w_{ij}(y_i - \overline{y})}{\frac{\sum_{i=1}^{n}(y_i - \overline{y})^2}{n}}. \tag{3.16}$$

Jedoch können die Daten eine räumliche Autokorrelation aufweisen, obwohl sie nicht existent ist. Ein möglicher Grund dafür ist eine Unterspezifikation, oder eine falsch getroffene Entscheidung bezüglich der Wahl der Gewichtungsmatrix, die mit den eigentlichen Beziehungen zwischen den räumlichen Einheiten nicht übereinstimmt.

3.3.3 Räumliche Regression

Da die Standardannahme der KQ-Schätzung bezüglich der Unabhängigkeit der Beobachtungen im Fall einer räumlichen Autokorrelation verletzt ist, versucht man die Daten entweder räumlich zu modellieren, oder die Autokorrelation, durch Veränderungen in dem ursprünglichen Modell zu vermeiden. Dieser Abschnitt konzentriert sich dennoch auf die räumliche Regression, insbesondere auf die simultanen autoregressiven Modelle.

3.3.4 Simultanes Autoregressives Modell

Die Spezifikation der SAR benutzt eine Regression der Werten der anderen räumlichen Einheiten, um die räumliche Abhängigkeit zu erfassen. Das heißt, dass die Fehlerterme ε so modelliert sind, dass sie voneinander wie folgt abhängig sind:

$$e_i = \sum_{i=1}^{m} b_{ij} e_i + \varepsilon_i. \tag{3.17}$$

Die Fehlerterme ε entsprechen den Residuen, die laut den Annahmen unter der Normalverteilung mit dem Mittelwert null und einer Diagonalmatrix der Kovarianzen $\sum_{\varepsilon}$ mit den Elementen $\sigma^2_{\varepsilon_i}$, wobei $i = 1, \ldots, m$, unabhängig verteilt sind. Die Werte von b_{ij} entsprechen der räumlichen Abhängigkeit zwischen den Regionen und die Werte b_{ii} werden gleich null gesetzt, damit keine Region mit sich selbst regressiert wird. Werden die Fehler als $e = B(Y - X'\beta) + \varepsilon$ dargestellt, kann man das Modell folgendermaßen ausgedrücken:

$$Y = X'\beta + B(Y - X'\beta) + \varepsilon. \tag{3.18}$$

Weiterhin kann das Modell in der Martrixnotation formuliert werden:

$$(I - B)(Y - X'\beta) + = \varepsilon, \tag{3.19}$$

wobei B eine Matrix der Abhängigkeitsparameter b_{ij} und I die Einheitsmatrix ist. Die Matrix $I - B$ muss regulär sein. Oft wird das Modell räparametrisiert, indem man die Matrix B wie folgt beschreibt: $B = \lambda W$. λ umfasst den Parameter der räumlichen Autokorrelation, dabei ist W eine Matrix, die eine räumliche Abhängigkeit beschreibt (vgl. Bivard, Pebesma und Gómez-Rubio (2008)). Dieses Modell wird weiterhin mit der Maximum-Likelihood-Methode geschätzt.

3.4 Modellvergleich

Es wird gesagt, dass die Probleme, die bei der Anpassung eines räumlichen Modells mit einer räumlichen Autokorrelation unter Umständen auftreten, sehr anspruchsvoll sein können. Manchmal kann diese Autokorrelation durch eine Unterspezifikation des Modells verursacht sein. In einem solchen Fall sollte man sich eher auf das ursprüngliche Modell konzentrieren, um die Probleme zu beheben. Wenn das räumliche Modell jedoch gut angepasst werden konnte, sollte eine Entscheidung bzgl. der Wahl des Modells getroffen werden. Dabei werden das Bestimmtheitsmaß, das Pseudo-Bestimmtheitsmaß und das Akaike Informationskriterium berücksichtigt.

Das Bestimmtheitsmaß R^2 gibt an, welcher Anteil der Varianz des Modells durch das Regressionsmodell erklärt wird. Dies kann folgendermaßen beschrieben werden:

$$R^2 = \frac{\frac{1}{n}\sum(\hat{y}_i - \overline{y})^2}{\frac{1}{n}\sum(y_i - \overline{y})^2} = 1 - \frac{\frac{1}{n}\sum(\hat{\varepsilon}_i - \overline{\hat{\varepsilon}})^2}{\frac{1}{n}\sum(y_i - \overline{y})^2}. \tag{3.20}$$

Dabei ist zu beachten, dass R^2 nur Werte zwischen null und eins annimmt, je größer der Wert, desto bessere Anpassung gefunden wurde. R^2 lässt sich jedoch nicht für Modelle nach der ML-Schätzung bestimmen. Deswegen ist ein Konzept von Pseudo-Bestimmtheitsmaßen entstanden, bei dem die verschiedenen Interpretationen von R^2, sowie die erklärte Varianz und das Quadrat der Korrelation, genügen. Die Werte der Pseudo-Bestimmtheismaße liegen ebenso zwischen null und eins, wobei die höheren Werte eine bessere Anpassung bedeuten. Eine der Möglichkeiten das Pseudo-Bestimmtheismaß

zu ermittelt, wurde von Nagelkerke bestimmt und lässt sich folgendermaßen beschreiben:

$$R^2_{Nagelkerke} = \frac{1 - \left(\frac{L_0}{L_1}\right)^{2/n}}{1 - L_0^{2/n}}. \tag{3.21}$$

Nach Nagelkerkes Bedingungen sollte Pseudo-Bestimmtheitsmaß mit R^2 übereinstimmen, wenn beide berechnet werden können.

Das Akaike Informationskriterium misst die relative Güte eines Modells und gibt einen Bias-Varianz-Trade-Off an. Für die KQ-Schätzung kann die Formel wie folgt beschrieben werden:

$$AIC := \ln\left(\frac{1}{N}RSS\right) + \frac{2K}{N} = \ln\left(\frac{1}{N}\sum_{i=1}^{N} e_i^2\right) + \frac{2K}{N} \tag{3.22}$$

wobei e_i die KQ-Residuen sind, N die Anzahl der Beobachtungen und K die Anzahl der Regressoren ist. Diese Gleichung zeigt, dass der zu erwartende Bias umso geringer ist, desto komplexer ein Modell ist, aber die Variabilität der Schätzungen höher ausfällt. Im Fall der ML-Schätzung kann die AIC mit Hilfe einer anderen Formel betimmt werden:

$$AIC = 2k - 2\ln(L) \tag{3.23}$$

Die Faustregel besagt, dass das Modell mit dem kleinsten AIC-Wert gewählt werden sollte.

4 Deskriptive Statistik

In diesem Kapitel werden die Variablen aus dem vorbereiteten Datensatz analysiert, um die Wohnungscharakteristiken in den einzelnen Ortsteilen nahezubringen. Sie werden sowohl graphisch auf den Karten dargestellt, als auch in den Tabellen beschrieben. Die Karten, die zur Veranschaulichung der deskriptiven Statistiken gedient haben, sind mithilfe der spdep Bibliothek der Programmiersprache R und der Shapefile-Datei entstanden.

4.1 Analyse der Wohnungscharakteristiken

Die Anzahl der Beobachtungen für die Ortsteile aus dem Datensatz zu Immobilien ist im Anhang in der Tabelle 10 zu finden. Da vor allem mit den Durchschnittswerten gearbeitet wurde, ist es wichtig, dass genug Beobachtungen betrachtet werden können. In vier Ortsteilen aus dem Datensatz wurden innerhalb von zehn Jahren weniger als fünf Immobilien verkauft, und zwar in der Stadtrandsiedlung Malchow, in Blankenfelde, Falkenber und Malchow.In letzterem wurde sogar nur ein Kaufvertrag abgeschlossen. Deswegen sollten die deskriptiven Statistiken zu diesen Ortsteilen mit Zurückhaltung interpretiert werden.

Die Variable, die im Mittelpunkt der Analysen steht, ist der Kaufpreismedian je m^2 in €. Sie wird die Zielgröße in der Regressionsanalyse sein. Zu dieser Variable gibt es keine fehlenden Werte und somit bestehen 96 Beobachtungen. Es fällt gleich auf, dass die Spannweite sehr breit ausfällt und 2392.5€ beträgt, was aus der Tabelle 2 hervorgeht. Es ist außerdem zu beobabchten, dass der Median und der Mittelwert sehr nah aneinander liegen, was eine symmetrische Verteilung vermuten lässt.

Variable	Obs	Mean	Median	Std. Dev.	Min	Max
KP	96	1465	1467	416.74	464	2856.5

Tabelle 2: Deskriptive Statistiken zu dem Kaufpreismedian je m^2 in €.

Die niedrigsten Werte befinden sich vor allem im Bezirk Marzahn-Hellersdorf, mit dem Minimum in Hellersdorf mit 464€ je m^2 und in Neukölln. Die größten Werte, wie in der Abbildung 2 zu sehen ist, kann man in Mitte (das Maximum im Tiergarten mit 2856.5€ je m^2) und im Steglitz-Zehlendorf auffinden. Die Skalierung der Legende wurde nach dem Jenks-Caspall-Algorithmus skaliert. Dieser Algorithmus zur Klassifikation von Werten minimiert die Unterschiede innerhalb einer Klasse und gleichzeitig maximiert er die Unterschiede zwischen den Klassen. Somit sollen sich die Ortsteile aus der selben Klasse in deren Kaufpreismedianen je m^2 am meisten ähneln[9].

[9] Außer der Grafiklegende zum Kaufpreismedian je m^2 wurden die Legenden mittels Quantilen bestimmt.

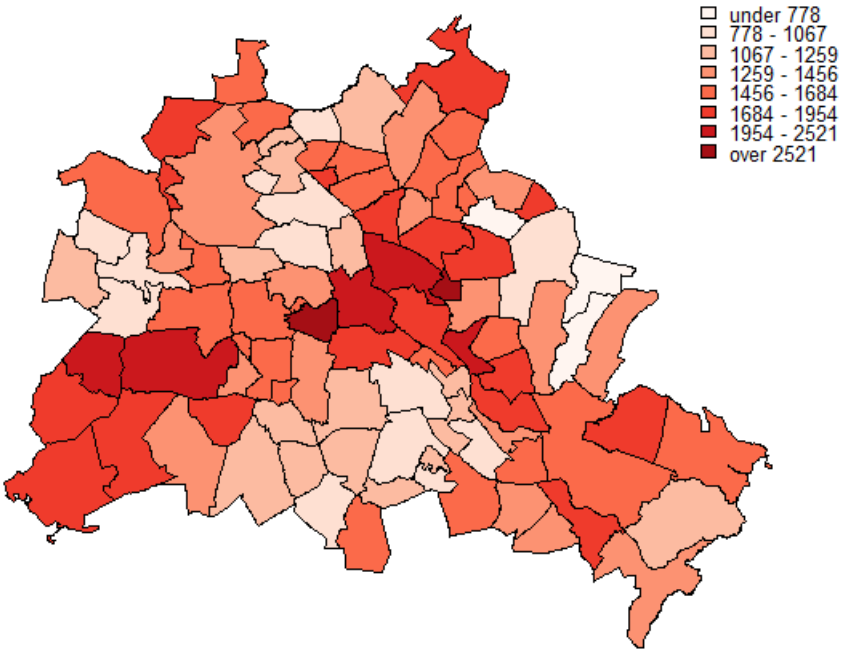


Abbildung 2: Kaufpreismedian je m^2 in €.

Die weiteren Variablen beziehen sich auf die Immobiliencharakteristiken in den jeweiligen Ortsteilen, sowie auf das Baujahr (hier als Anteil der Altbauten an den verkauften Immobilien), das Stockwerk, in dem die Wohnungen durchschnittlich liegen, die durchschnittliche Raumanzahl, die durchschnittliche stadträumliche Wohnlage und die durchschnittliche Wohnfläche der Immobilien in demselben Ortsteil.

Variable	Obs	Mean	Median	Std. Dev.	Min	Max
Baujahr	96	35.51%	33.45%	28.41%	0%	100%
Ge	94	2.31	2.23	0.79	1.00	6.02
Ra	96	2.82	2.74	0.43	2.06	5.00
SW	96	4.74	4.57	1.38	3.00	8.66
WF	96	79.18	78.00	12.21	53.74	109.00

Tabelle 3: Deskriptive Statistiken zu den Wohnungscharakteristiken.

Außer des durchschnittlichen Stockwerks, mit fehlenden Werten für Malchow und Stadtrandsiedlung Malchow, sind die Variablen du den Immobiliencharakteristiken vollständig und enthalten 96 Beobachtungen. Es ist deutlich zu sehen, dass es große Unterschiede zwischen den Altbauanteilen an verkauften Immobilien in den Berliner Ortsteilen gibt. Dies wird in der Abbildung 3 veranschaulicht. In Lichtenberg liegt er zum Großteil bei null, Falkenberg, Malchow, Neu-Hohenschönhausen und Rummelsburg, aber auch für Charlottenburg-Nord, Gropiusstadt, Müggelheim und Hellersdorf. In der Stadtrandsiedlung Malchow sind alle verkauften Wohnungen vor 1949 entstanden.

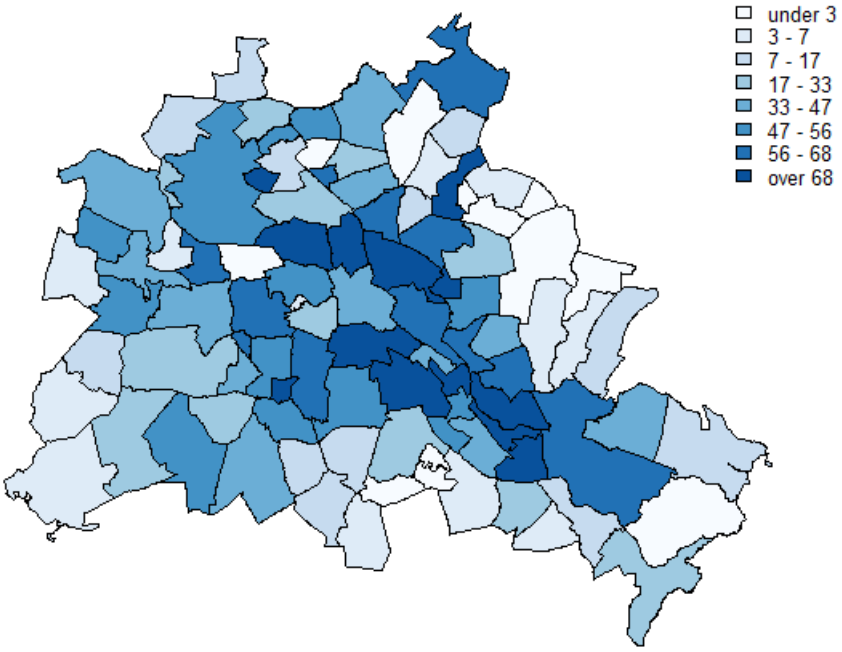


Abbildung 3: Anteil der Altbauten an allen verkauften Immobilien in Prozent.

Zu großen Dispersionen, mit einem Minimum - ein Stockwerk - in Gatow und im Märkischem Viertel, kommt es zwischen den durchschnittlichen Stockwercken, der jeweiligen Ortsteile. Die Wohnungen, die in den durchschnittlich höchsten Stockwercken liegen, werden in Hansaviertel verkauft. Aus dem Datensatz geht hervor, dass die verkauften Wohnungen mit der durchschnittlich kleinsten Anzahl der Räume in Gesundbrunnen liegen und jene mit der größten in Malchow (mit durchschnittlich fünf Räumen). Alle Wohnungen, die in Gesundbrunnen, Märkischem Viertel und Borsigwalde verkauft wurden, befinden sich in einer einfachen Wohnlage. Dagegen sind die Wohnungen in Grunewald, solche mit der sehr guten Wohnlage. Weiterhin kann man feststellen, dass sich die Wohnungen mit der kleinsten Wohnfläche in Siemensstadt befinden. Die mit der größten Fläche sind in Malchow, Grunewald, Dahlem und Schmöckwitz zu finden.

4.2 Analyse der Nachbarschaftscharakteristiken

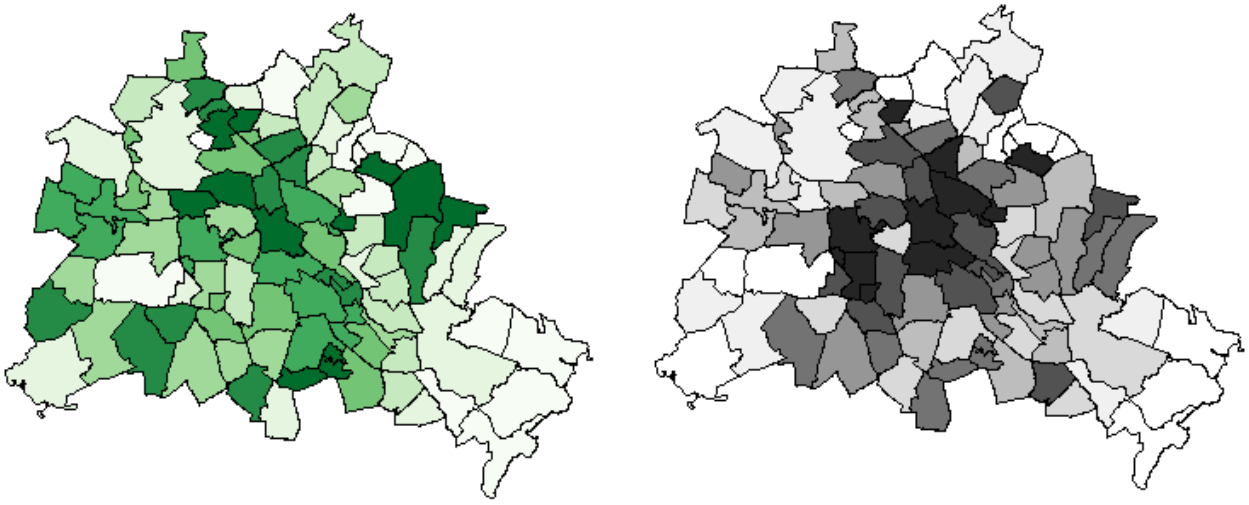
Die andere Gruppe der Variablen, die später bei der Erklärung des Kaufpreismedians je m^2 behilflich wird, sind die Nachbarschaftscharakteristiken.

Variable	Obs	Mean	Median	Std. Dev.	Min	Max
Park	95	5.89%	5.01%	5.02%	0.08%	23.93%
Strassen	95	12.87%	12.88%	7.69%	0.51%	35.62%

Tabelle 4: Deskriptive Statistiken zur Flächennutzung.

Zu den Variablen bezüglich des Anteils der Straßen und Parkflächen an der Gesamtflä-

che des Ortsteils sind 95 Beobachtungen verfügbar. In beiden Fällen ist der Wert für Borsigwalde fehlend.



(a) (b)

Abbildung 4: Anteil der Parkflächen (Abbildung 4a) und der Straßen (Abbildiung 4b) an der Gesamtfläche des Ortsteils.

In der Abbildung 4a kann man erkennen, dass sich die Ortsteile mit dem größten Park- und Grünflächenanteil vor allem in Mitte und Marzahn-Hellersdorf - mit einem Maximum von 23.93% in Neu-Hohenschönhausen (Lichtenberg) - befinden. Die Ortsteile mit dem kleinsten Grünflächenanteil sind in Treptow-Köpenick mit einem Minimum von 0.08% in Müggelheim konzentriert. Die Angaben zu Anteilen der Straßen an der Gesamtfläche sind keine Überraschung. Die größte Dichte ist im zentralen Teil Berlins zu finden, in Mitte, Pankow, Friedrichshain-Kreuzberg und Charlottenburg-Wilmersdorf. Dagegen ist die kleinste Straßendichte am Rand der Stadt gelegen, was die Abbildung 4b veranschaulicht.

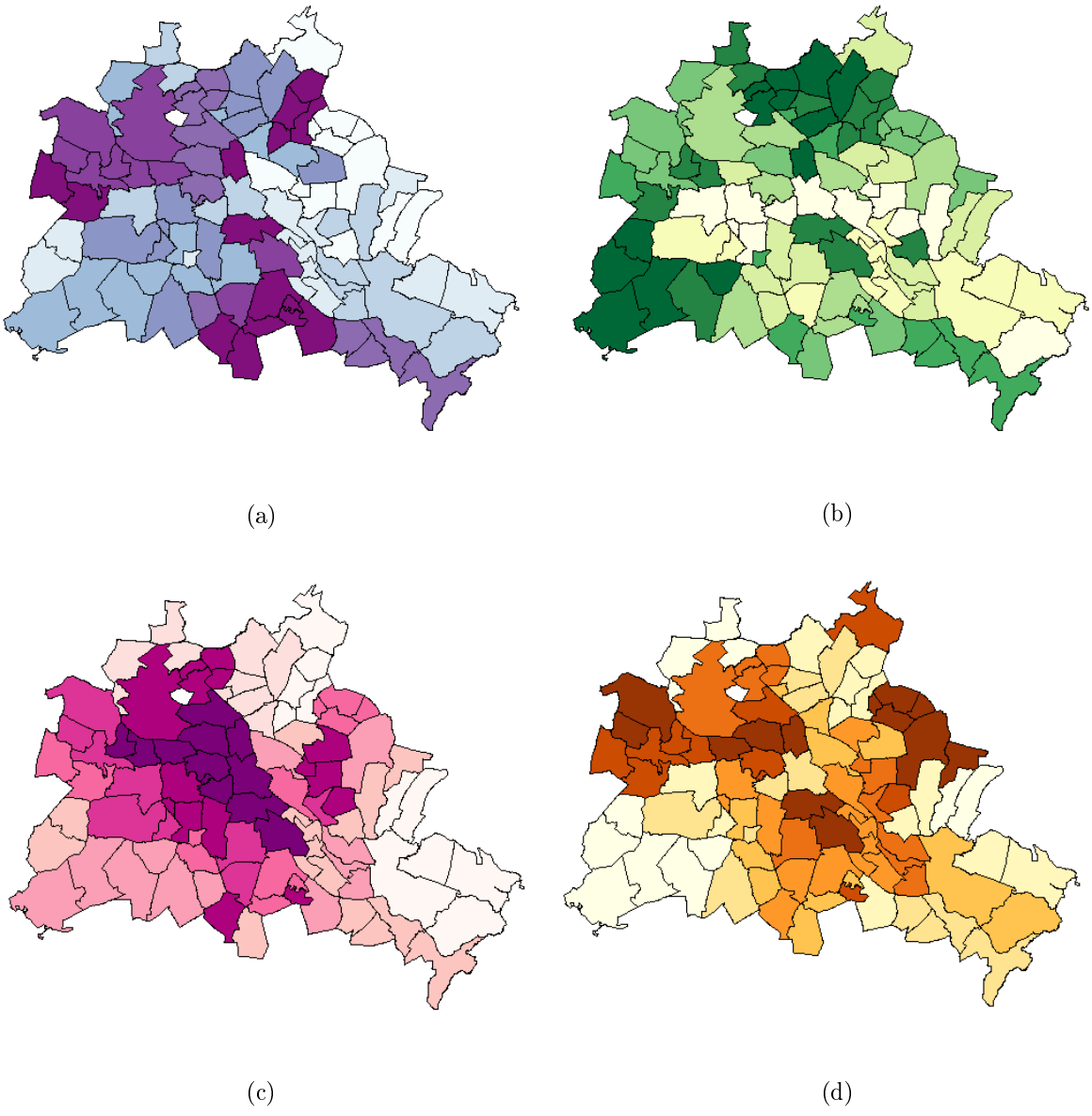
Zunächst werden die Angaben zu den sozio-ökonomischen Daten analysiert. Die Angaben beziehen sich wieder auf 95 Ortsteile, da Borsigwalde nicht berücksichtigt wurde.

Variable	Obs	Mean	Median	Std. Dev.	Min	Max
Straftaten	95	11330	9918	5767.54	4252	37440
Schulen	95	7.46	5.00	7.14	0.00	39.00
Sozialmietwohnungen	95	7.62%	4.70%	7.62%	0.00%	38.70%
unter18	95	14.94%	15.18%	1.79%	11.76%	19.30%
ausland	95	7.00%	6.33%	5.32%	0.71%	24.81%
arbeitslos	95	7.99%	7.71%	3.19%	2.69%	15.04%

Tabelle 5: Deskriptive Statistiken zu den sozio-ökonomischen Daten.

Es fällt auf, dass große Ungleichheiten in den Kriminalitätsraten der Berliner Ortsteile vorliegen und deren Verteilung rechtsschief ist. Ein hoher Wert der Standardabweichung

weist wieder auf deutliche Variabilität zwischen den Regionen hin. Die wenigsten Straftaten je 10.000 Einwohner weisen die Ortsteile am Rand Berlins auf, so wie Müggelheim und Mahlsdorf. Als die gefährlichsten Ortsteile können mit über 30000 Straftaten je 10.000 Einwohner Mitte und Tiergarten bezeichnet werden. Jedoch hat u.a. auch die touristische Attraktivität eines Ortsteils Einfluss auf die Häufigkeitszahl von Straftaten, die Touristen können sowohl Opfer, als auch Täter der Straftaten sein. Somit können sie die Regionalstatistiken im Zentrum Berlins erhöhen.
Die Anzahl der Schulen hängt ebenfalls stark von der Entfernung vom Stadtzentrum ab. Die Ortsteile, in denen sich keine Schulen befinden, liegen vor allem am Rand, sowie Stadtrand Siedlung Malchow, Wartenberg und Konradshöhe. Dabei ist Kreuzberg der Bezirk mit der mit Abstand größten Anzahl an Schulen.



(a) (b)

(c) (d)

Abbildung 5: Abbildungen zu den sozio-ökonomischen Daten, dem Anteil der Sozialmietwohnungen (a), dem Anteil der Population unter 18 (b), dem Anteil der Ausländer (c) und der Arbeitslosenquote (d).

Zu starken Ungleichmäßigkeiten kommt es auch in der Verteilung von Sozialmietwohnungen. Die Mehrheit solcher Wohnungen befindet sich in westlichen Teilen Berlins, vor allem in Spandau, Neukölln und Tempelhof-Schöneberg, mit einem Maximum von 38.7% in Marienfelde. Es gibt jedoch auch Ortsteile ohne Sozialmietwohnungen, die zum Großteil in Lichtenberg liegen.
Wie in der Abbildung 5b zu sehen ist, wohnt die Mehrheit der Minderjährigen in Pankow, Reinickendorf und Steglitz-Zehlendorf, mit einem Anteil von über 17%. Dennoch gibt es in Berlin Gebiete, in denen Kinder und Jugendliche weniger als 12% Population darstellen, wie in Teilen von Lichtenberg und Marzahn-Hellersdorf.
Die Tabelle 5 und die Abbildung 5c stellen den Anteil der Ausländer dar. Es fällt auf, dass Ausländer vor allem im Zentrum Berlins wohnen, in Wedding und Neukölln sind das knapp 25% der Einwohner. Ortsteile, die unter Ausländern wenig populär zu sein scheinen, sind die östlichen Teile Berlins, wie Marzahn-Hellersdorf, was historisch bedingt ist.
Die letzte untersuchte Variable ist die Arbeitslosenquote der jeweiligen Ortsteilen, welche im Durchschnitt bei knapp 8% liegt. Jedoch gibt es in Berlin Ortsteile mit unter 4% Arbeitslosigkeit. Diese sind vor allem Teile von Steglitz-Zehlendorf. In den Ortsteilen, in denen Arbeitslosigkeit ein existierendes Problem darstellt, nämlich in Wedding, Gesundbrunnen und Neukölln, liegt sie bei über 13%.

Variable	Obs	Mean	Median	Std. Dev.	Min	Max
Haltestellen	95	31.57	27.00	22.27	0.00	106.00
Ubahn	95	1.74	0.00	3.39	0	21.00
Sbahn	95	1.24	1.00	1.60	0.00	7.00

Tabelle 6: Deskriptive Statistiken zur Haltestellenanzahl.

Weitere Informationen, die der vorliegende Datensatz beinhaltet, beziehen sich auf die Anzahl der Haltestellen in den jeweiligen Ortsteilen. Hier ebenfalls existieren 95 Beobachtungen, zu allen Ortsteilen ausschließlich Borsigwalde. Aus der Tabelle 6 kann man entnehmen, dass große Unterschiede in der Verteilung der Haltestellen in Berlin bestehen. Es gibt Ortsteile, wie Malchow, ohne Haltestellen und welche, wie Mitte mit 106 Haltestellen. Die gleiche Tendenz ist - getrennt betrachtet - für Ubahn- und Sbahnstationen zu sehen.

5 Regressionsanalyse

In diesem Kapitel wird die Regressionsanalyse zur Erklärung des Kaufpreismedians je m^2 durchgeführt. Zuerst wird das additive Regressionsmodell erstellt und als nächstes das simmultane autoregressive Modell, um die räumliche Autokorrelation zu berücksichtigen. Die Ergebnisse der beiden Modelle werden analysiert und die Standardannahmen werden überprüft. Letztendlich wird eine Entscheidung getroffen, welches Modell die betrachtete Zielgröße besser erklärt.

5.1 Die allgemeine Regression

Zunächst wurde der Ortsteil Borsigwalde aus der Regressionsanalyse ausgeschlossen, da zu diesem, erst seit 2012 abgegrenzten Ortsteil, keine Angaben zu den Nachbarschaftscharakteristiken vorhanden sind. Die Variable die im Mittelpunkt der Arbeit steht, ist der Kaufpreismedian je m^2. Da vermutet wird, dass eine nicht-lineare Beziehung zwischen dem Preis und den Attributen besteht, wird mit der logarithmierten Form dieser Variable gearbeitet. Dies kann außerdem das Problem der Heteroskedastiztät mildern. Die Variable zur durchschnittlichen Raumanzahl wurde auch transformiert und weiterhin quadriert, um die abnehmenden Grenzerträge des Attributs zu verdeutlichen. Weiterhin wurden die Variablen bezüglich der Flächennutzung logarithmiert, ebenfalls um die nicht-lineare Beziehung zu beachten.

Nachdem die erste Regression mit allen Variablen durchgeführt wurde, wurde überprüft, ob sie von Ausreißern beeinflusst wird. Dies wurde mit der Cook's Distanz untersucht, was in der Abbildung 6 zu sehen ist. Die Beobachtungen, deren Nummer in der Abbildung zu sehen sind, haben besonders großen Einfluss auf die Regression, zu nennen sind das Hansaviertel, Kaulsdorf und Mahlsdorf. Die beiden letzteren zeichnen sich durch einen der niedrigsten Kaufpreiswerte. An dieser Stelle werden die drei Ortsteile aus der Regressionsanalyse ausgenommen.

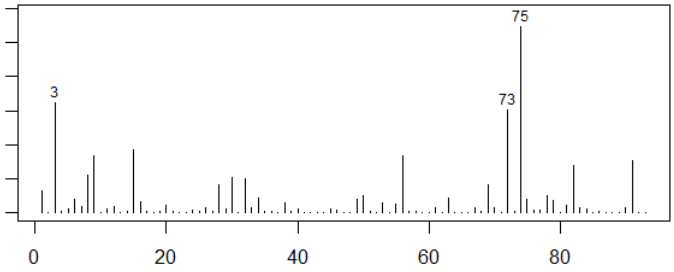


Abbildung 6: Cook's Distanz aller Beobachtungen der Regression.

Da sich nicht alle Variablen der Regression, die alle Angaben aus dem Datensatz beinhaltet, sich als signifikant herausgestellt haben, wurde die Funktion step in R benutzt. Diese Funktion wählt die Komponente des Regressionsmodells so, dass das Akaike Informationskiterium möglichst klein gehalten wird. Das empfohlene Modell sieht wie folgt aus:

$$\begin{aligned}\log(KP) = \beta_0 + \beta_1 Ge + \beta_3 Ra^2 + \beta_4 WF + \beta_5 \log(Strassen) + \beta_6 \log(Park) \\ + \beta_7 Straftaten + \beta_8 Sozialmietwohnungen + \varepsilon\end{aligned} \tag{5.1}$$

In der Tabelle 7 werden die Ergebnisse der Regression aus der Gleichung 5.1 dargestellt. Der Achsenabschnitt und die durchschnittliche Wohnfläche sind auf jedem üblichen Signifikanzniveau signifikant. Das durchschnittliche Stockwerk, die durchschnittliche quadrierte Raumanzahl, der logarithmierte Anteil der Straßen, die Häufigkeitszahl der Straftaten und Anteil der Sozialmietwohnungen können mit einer Irrtumswahrscheinlichkeit von 5% als signifikant betrachtet werden. Der logarithmierte Anteil der Park- und Grünflächen hat auf einem Signifikanzniveau von 10% einen bedeutenden Einfluss auf die Zielgröße und den Kaufpreismedian je m^2.

Coefficients	Estimate	Std. Error	t value	Pr(>|t|)
Intercept	6.85	0.79	27.22	$< 2e^{-16}$
Ge	-1.20	0.50	-2.388	0.019
Ra^2	-1.29	0.41	-3.139	0.002
WF	0.27	0.16	6.01	0.0002
log(Strassen)	1.02	0.45	2.28	0.025
log(Park)	-0.68	0.37	-1.81	0.075
Straftaten	0.007	0.01	2.23	0.029
Sozialmietwohnungen	-3.00	1.00	-2.99	0.004

Tabelle 7: Regressionsergebnisse.

Der logarithmierte Wert des Achsenabschnittes gibt den Fixpreis einer beliebigen Wohnung in Berlin an. Nach einer Umrechnung mit der Exponentialfunktion bekommt man einen Wert von 943.88 € je m^2. Einen negativen Effekt auf den Kaufpreis haben der logarithmierte Anteil der Parkflächen, dessen Erhöhung um 1% einen Abfall des Kaufpreises um 0.68% verursacht, und der Anteil der Sozialwohnungen, dessen Anstieg um 1% einem Abfall des Kaufpreises um 3% einhergeht. Dieses Ergebnis ist im Fall der Sozialmietwohnungen nicht überraschend. Für die Parkflächen hingegen ist dies ein durchaus überraschender Effekt. Dies kann dadurch verursacht werden, dass sich die großen Grünflächen am Rand Berlins befinden, so wie in Marzahn und Buckow (siehe Abbildung 4a), wo die Preise relativ niedrig sind. Einen stärkeren negativen Effekt hat das durchschnittliche Stockwerk. Wenn die gleiche Wohnung im selben Gebäude, ein Stockwerk höher liegt, hat sie einen um 1.2% niedrigeren Preis. Auch der logarithmierte Anteil der Straßen hat einen bemerkenswerten Einfluss auf die Zielgröße hat , dessen Erhöhung um 1% bewirkt einen Anstieg des Preises um 1.02% bewirkt.

5.2 Die räumliche Regression

Es wird nun versucht, das räumliche Modell auf die vorher benutzten Daten anzuwenden. Zuerst wurden mithilfe der spdep Bibliotkek, die zur die Durchführung von räumlichen Analysen entwickelt wurde, wurden die Nachbarschaften bestimmt. Dazu wurde die Methode der vier nächsten Nachbarn gewählt, das heißt für jeden Orteil eine Liste dessen vier nähsten Nachbarn erstellt wird. Weiterhin werden die Gewichtungen so bestimmt, dass sie sich für jeden Ortsteil zu eins summieren (siehe Kapitel 3.3.1).

Das räumliche Modell berücksichtigt die möglichen Autokorrelationen zwischen den Regionen. Um zu überprüfen, ob im Fall der vorliegenden Daten tatsächlich eine räumliche

Beziehung vorliegt, wurde Moran's I verwendet. Dazu wird die Funktion lm.morantest benutzt, die einen Moran's I Test auf Autokorrelation in den Residuen des vorherigen Regressionsmodells auf Basis der vorher bestimmten Gewichtungen vornimmt. Der beobachtete Moran's I Wert beträgt 0.09 auf dem Signifikanzniveau von 5%. Dies ist zwar kein Zeichen für eine beträchtliche räumliche Autokorrelation, aber es stellt auch keinen Grund dar, auf das räumliche Modell zu verzichten. In der Abbildung 7 wurden die Werte für den lokalen Moran's I veranschaulicht. Es fällt auf, dass die stärkste Autokorrelation für die Ortsteile im Zentrum Berlins, Steglitz-Zehlendorf, Marzahn-Hellersdorf und Neukölln, besteht.

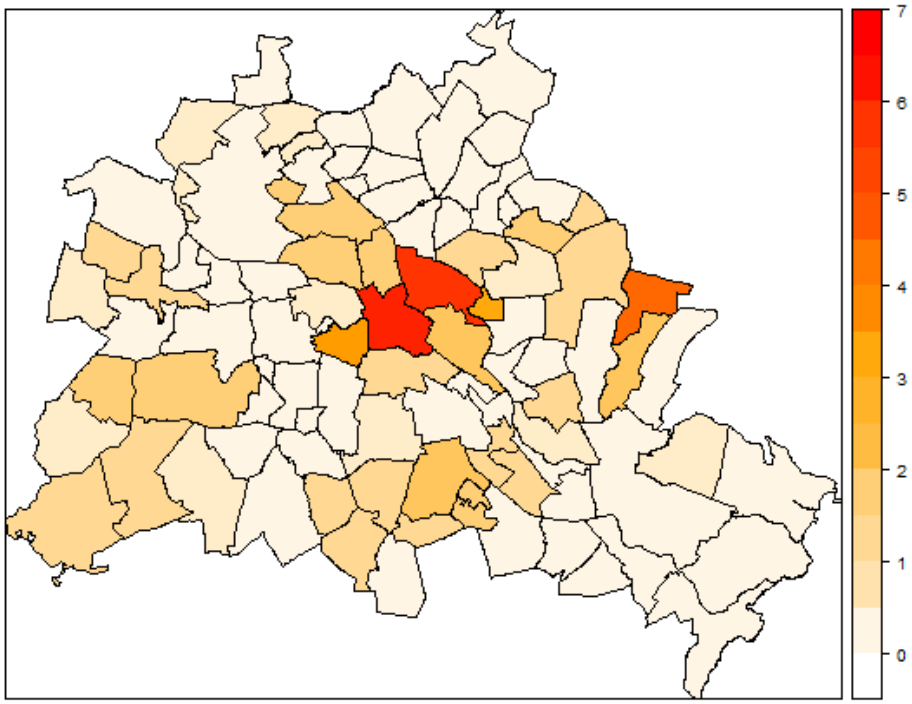


Abbildung 7: Lokal Moran's I für Berliner Ortsteile.

An dieser Stelle wird ein simultanes autoregressives Modell auf Basis des vorher erstellten additiven Regressionsmodells und der Gewichtungen angefertigt. Dieses wird mithilfe der spautolm Funktion, die auf dem simultanen autoregressiven Modell und der Maximum-Likelihood-Schätzung basiert, zustande gebracht. Die Ergebnisse dieser Regression sind in der Tabelle 8 dargestellt.

Die Ergebnisse aus dem simultanen autoregressiven Modell unterscheiden sich hauptsächlich nicht stark von denen aus dem additiven Modell (vgl. Tabelle 7). Die Variablen sind mit den selben Irrtumswahrscheinlichkeiten ähnlich signifikant, die Werte der t-Statistik unterscheiden sich minimal. Die Vorzeichen bei den Koeffizienten bleiben unverändert und die Koeffizienten selbst ändern sich kaum. Der Fixpreis für eine beliebige Wohnung in Berlin liegt in diesem Modell ein wenig höher, bei 962,95€.

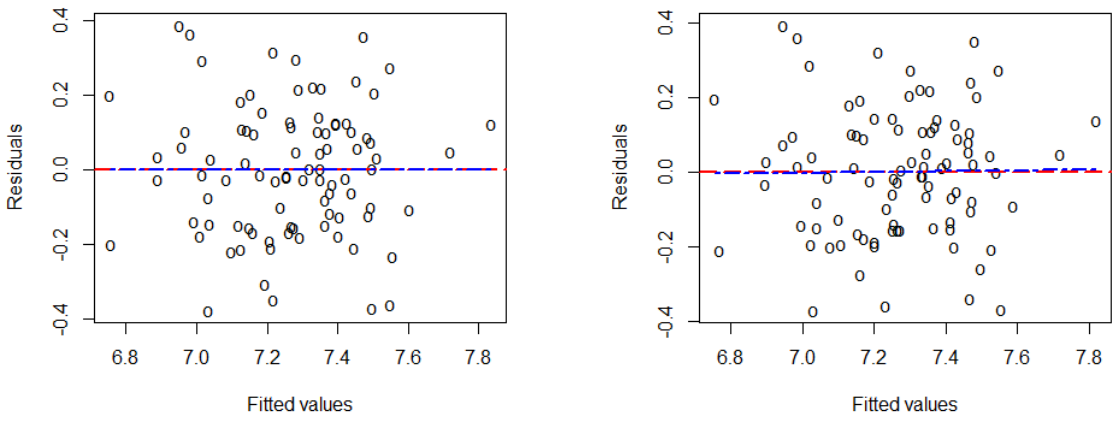
Coefficients	Estimate	Std. Error	t value	Pr(>|t|)
Intercept	6.87	0.89	28.26	$< 2e^{-16}$
Ge	-1.25	0.48	-2.57	0.01
Ra^2	-1.26	0.39	-3.26	0.001
WF	0.26	0.15	6.17	0.0003
log(Strassen)	0.96	0.45	2.09	0.037
log(Park)	-0.67	0.38	-1.77	0.076
Straftaten	0.007	0.01	2.26	0.024
Sozialmietwohnungen	-2.91	0.99	-2.93	0.003

Tabelle 8: Regressionsergebnisse der räumlichen Regression.

5.3 Modellwahl und Diagnostik

Zunächst werden die Standardannahmen der Modelle überprüft. Es wird auf Homoskedastizität, Normalverteilung der Residuen und mögliche Multikollinearität geachtet. Nachdem dies kontrolliert wurde, werden die zwei Modelle hinsichtlich deren Bestimmtheitsmaßen und Akaikes Informationskriteriums verglichen.

Um auf das Vorliegen einer Multikollinearität zu testen, wurden die Werte für den Varianzinflationsfaktor ausgerechnet. Der höchste Wert von 4.45 ergibt sich dabei für die Raumanzahl. Für die Wohnfläche beträgt der Wert 3.96. Dies impliziert eine hohe Korrelation zwischen den beiden Variablen. Für die restlichen Variablen sind die Werte nicht größer als 2.5. Da die Varianzinflationsfaktoren aller Variablen fünf nicht überschreiten, wird angenommen, dass Multikollinearität kein beträchtliches Problem darstellt.



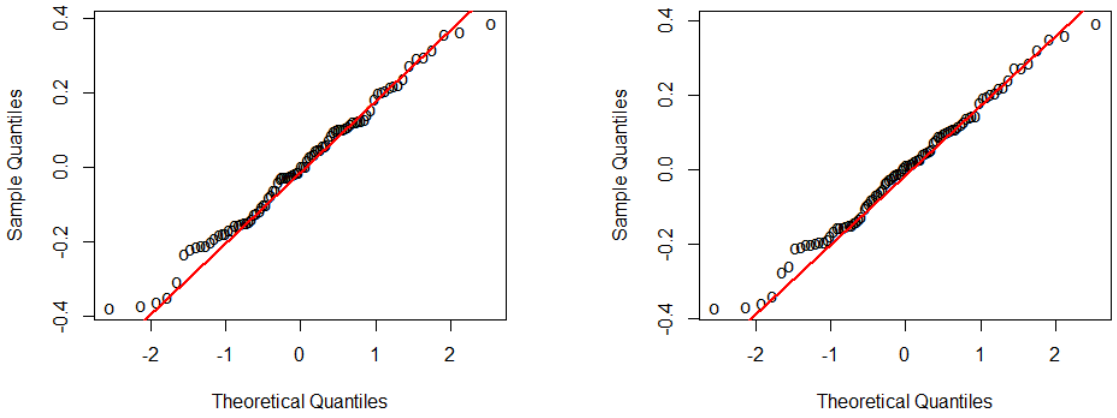
(a) (b)

Abbildung 8: Die Abbildung der angepassten Werte und Residuen des additiven Regressionsmodells (8a) und des simultanen autoregressiven Modells (8b).

Die Annahme der Homoskedastizität der Residuen kann grafisch anhand der Abbildung 8 nachgeprüft werden. Da keine der beiden Grafiken eine für die Heteroskedastizität übliche

Trichterform annimmt, kann angenommen werden, dass die Fehler homoskedastisch sind und die Annahme somit erfüllt ist. Der für das additive Modell durchgefürte Breusch-Pagan-Test bestätigt die Aussage, dass die Hypothese der Homoskedastizität nicht verworfen werden kann.

Eine weitere Standardannahme der Regressionsanalyse ist die Normalität der Fehlerterme. Dies kann nach Betrachtung der QQ-Plots, die in der Abbildung 9 veranschaulicht wurden, festgestellt werden. Die beiden unterscheiden sich nicht stark voneinander. Es bestehen Abweichungen an den Enden der Verteilungen, jedoch kann vermutet werden, dass die Annahme nicht verletzt ist. Um eine sichere Aussage darüber zu treffen, wird der Shapiro-Wilk-Test auf Normalverteilung durchgeführt. Mit beiden p-Werten weit über 0.05 kann die Hypothese der Normalverteilung nicht abgelehnt werden.



(a) (b)

Abbildung 9: QQ-Plots der des additiven Regressionsmodells (9a) und des simultanen autoregressiven Modell (9b).

Es kann nun festgehalten werden, dass die beiden Modelle die Standardannahmen erfüllen. In beiden Modellen wurde eine Homoskedastizität und Normalverteilung angenommen. An dieser Stelle werden die beiden Modelle verglichen, um die Entscheidung zu treffen, welches eine bessere Anpassung für die vorliegenden Daten darstellt.

Als erstes werden die Werte der Bestimmtheismaße betrachtet. Für das additive Regressionsmodell bekommt man einen R^2 von 57.42% und einen adjustierten R^2 von 53.78%. Laut des Nagelkerke Pseudo-Bestimmtheitsmaßes erklärt das autoregressive Modell 58.25%. Somit bildet das räumliche Modell eine bessere Anpassung.

	AM	SAR
AIC	-41.42	-41.19

Tabelle 9: Werte des Akaike Informationskriteriums.

Zunächst wird das Akaike Informationskriterium untersucht. Die dazugehörigen Werte kann man in der Tabelle 9 finden. Das ein wenig kleinere Ergebnis ist auf das additive

Modell zurückzuweisen. Folglich ist das additive Modell laut diesem Modellwahlkriterium vorteilhafter.

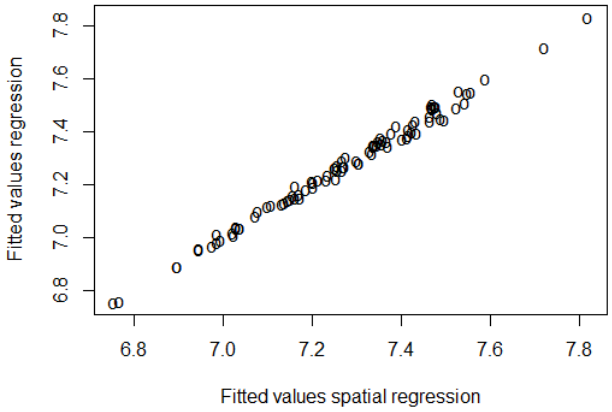


Abbildung 10: Grafik der angepassten Werte der beiden Modelle.

Da die beiden Kriterien unterschiedliche Entscheidungshinweise für die Modellwahl geliefert haben, werden jetzt die angepassten Werte aus den beiden Regressionen betrachtet, die in der Abbildung 10 abgebildet wurden. Aus dieser Graphik geht hervor, dass keine wesentlichen Unterschiede zwischen den beiden Modellen bestehen.

6 Fazit

Das Ziel dieser Bachelorarbeit war, anhand der gesammelten Daten die Einflussfaktoren des Kaufpreismedians je m^2 für die Berliner Ortsteile zu ermitteln. Dabei haben sich die Variablen - das durchschnittliche Stockwerk, die quadrierte durchschnittliche Raumanzahl, die durchschnittliche Wohnfläche, der logarithmierte Anteil der Straßen an der Gesamtfläche, der logarithmierte Anteil der Park- und Grünflächen an der Gesamtfläche (Signifikant auf dem Signifikanzniveau von 10%), die Häufigkeitszahl der Straftaten und Sozialmietwohnungenanteil an der Gesamtfläche - als signifikant erwiesen. Alle Variablen, außer des logarithmierten Anteils der Parkflächen und der Häufigkeitszahl der Straftaten, zeigen das erwartete Vorzeichen auf. Der leicht positive Wert des Koeffizienten der Häufigkeitszahl ist damit verbunden, dass die Ortsteile, die sich durch viel Tourismus auszeichnen, eine höhere Häufigkeitszahl der Straftaten aufweisen. Dabei sind diese Ortsteile aufgrund ihrer Attraktivität durch höhere Preise gekennzeichnet. Weiter liegen die Ortsteile mit einem großen Parkflächennteil zum Großteil in einer gewissen Entfernung zum Stadtzentrum, was das negative Vorzeichen beim logarithmierten Anteil der Parkfläche verursachen könnte.

Eine Verbesserung der Analyse der Berliner Immobilienpreise wäre möglich, wenn man mit den Koordinaten der Immobilien arbeiten könnte. Mithilfe dieser Koordinaten könnten die Distanzen zum nächsten Park, zum Stadtzentrum usw. errechnet werden. Diese Distanzen könnten letztendlich als Kovariaten in der Regression benutzt werden. Ein solcher Ansatz wurde u.a. von Martins-Filho und Bin für Immobilienpreise in Oregon angewendet. Außerdem könnte dieses Modell um Umweltcharakteristiken bereichert werden, insbesondere um Verkehrslärm (vgl. M.Theebe) und die Luftqualität (vgl. D. Harrison und D. Rubinfeld). Weiterhin wäre sinnvoll, die Variable bezüglich der Schulanzahl mit einem Schüler-Lehrer-Quotienten zu ersetzen, wie in ähnlichen Analysen geschehen (vgl. D. Harrison und D. Rubinfeld).

Zur Erklärung der Variabilität des Kaufpreismedians je m^2 wurden zwei Modelle erstellt: Das additive Regressionsmodell und das räumliche Regressionsmodell. Das Problem bei der Regressionsanalyse war, dass nicht genug Beobachtungen für jeden Datensatz zu den Immobiliencharakteristiken vorlagen, da in vier Ortsteilen innerhalb von zehn Jahren weniger als fünf Immobilien verkauft wurden. Außerdem wurden aus der Regressionsanalyse drei Ortsteile wegen deren besonders starken Einflusses ausgeschlossen. Dies waren das Hansaviertel in Mitte, sowie Kaulsdorf und Mahlsdorf in Marzahn-Hellersdorf.

Die beiden Regressionsmodelle haben ähnliche Ergebnisse geliefert. Eine leichte räumliche Autokorrelation zwischen den Ortsteilen wurde bestätigt. Sowohl das additive Modell, als auch das räumliche, erklären ca. 55%-60% der Variabilität des Kaufpreismedians zwischen den Ortsteilen. Dieses Ergebniss ist vergleichbar, mit denen aus anderen Studien zu ähnlichen Themen (vgl. H. W. Richardson, J. Vipond und R. A. Furbey). Eine Verbesserung dieses Ergebnisses wäre durch eine andere Wahl der Variablen (s.o.) plausibel.

In dieser Arbeit wurde angedeutet, dass in der Literatur empfohlen wird, sich auf das Modell ohne Berücksichtigung der räumlichen Autokorrelation zu konzentrieren. Jedoch bieten sich in diesem Fall einige Ansätze an, die man anwenden könnte, um das räumliche Modell zu verbessern. Erstens wäre das, eine andere Wahl des Nachbarschaftskriteriums, wie eine andere Anzahl der k-nächsten Nachbarn oder ein Kriterium, das auf einem Graphen basiert, zu treffen. Zweitens kann man sich für eine andere Art der Gewichtungsmatrix

entscheiden, z.B. eine, die annimmt, dass sich die Stärke der Nachbarschaftsbeziehungen mit steigender Distanz abschwächt. Auch könnten Ergebnisse aus einem anderen räumlichen Regressionsmodell analysiert werden, insbesodere das kondizionale autoregressive Modell zu erwähnen.

Wie in der Einleitung angedeutet, ist das Wohnungsangebot einer der wichtigsten Einflussfaktoren auf die Immobilienpreise. Dieser könnte jedoch in dieser Arbeit nicht analysiert werden. Um ein volles Bild des Berliner Immobilienmarktes zu bekommen, sollte eine Zeitreihenanalyse durchgeführt werden. Dieser Ansatz ermöglicht eine Betrachtung von den Trends, die auch von der Wohnungsangebot beeinflusst werden können.

A Anhang

Tabelle 10: Liste der Berliner Ortsteile

Nummer	Bezirk	Ortsteil	Anahl der Beobachtungen
101	Mitte	Mitte	8043
102	Mitte	Moabit	3472
103	Mitte	Hansaviertel	264
104	Mitte	Tiergarten	748
105	Mitte	Wedding	3017
106	Mitte	Gesundbrunnen	546
201	Freidrichshain-Kreuzberg	Friedrichshain	8122
202	Freidrichshain-Kreuzberg	Kreuzberg	5247
301	Pankow	Prenzlauer Berg	11183
302	Pankow	Weißensee	1379
303	Pankow	Blankenburg	44
304	Pankow	Heinersdorf	59
305	Pankow	Karow	89
306	Pankow	Stadtrandsiedlung Malchow	2
307	Pankow	Pankow	2062
308	Pankow	Blankenfelde	3
309	Pankow	Buch	313
310	Pankow	Französisch Buchholz	298
311	Pankow	Niederschönhausen	1308
312	Pankow	Rosenthal	396
313	Pankow	Wilhelmsruh	67
401	Charlottenburg-Wilmersdorf	Charlottenburg	8432
402	Charlottenburg-Wilmersdorf	Wilmersdorf	8053
403	Charlottenburg-Wilmersdorf	Schmargendorf	1949
404	Charlottenburg-Wilmersdorf	Grunewald	936
405	Charlottenburg-Wilmersdorf	Westend	2603
406	Charlottenburg-Wilmersdorf	Charlottenburg-Nord	426
407	Charlottenburg-Wilmersdorf	Halensee	1011
501	Spandau	Spandau	1487
502	Spandau	Haselhorst	693
503	Spandau	Siemensstadt	484
504	Spandau	Staaken	516
505	Spandau	Gatow	60
506	Spandau	Kladow	295
507	Spandau	Hakenfelde	925
508	Spandau	Falkenhagener Feld	564
509	Spandau	Wilhelmstadt	1496
601	Steglitz-Zehlendorf	Steglitz	3006
602	Steglitz-Zehlendorf	Lichterfelde	2167
603	Steglitz-Zehlendorf	Lankwitz	984
604	Steglitz-Zehlendorf	Zehlendorf	1971
605	Steglitz-Zehlendorf	Dahlem	410
606	Steglitz-Zehlendorf	Nikolassee	400

Tabelle 10: Liste der Berliner Ortsteile

Nummer	Bezirk	Ortsteil	Anahl der Beobachtungen
607	Steglitz-Zehlendorf	Wannsee	300
701	Tempelhof-Schöneberg	Schöneberg	7822
702	Tempelhof-Schöneberg	Friedenau	1773
703	Tempelhof-Schöneberg	Tempelhof	2697
704	Tempelhof-Schöneberg	Mariendorf	1910
705	Tempelhof-Schöneberg	Marienfelde	640
706	Tempelhof-Schöneberg	Lichtenrade	1504
801	Neukölln	Neukölln	4299
802	Neukölln	Britz	659
803	Neukölln	Buckow	1108
804	Neukölln	Rudow	979
805	Neukölln	Gropiusstadt	884
901	Treptow-Köpenick	Alt-Treptow	287
902	Treptow-Köpenick	Plänterwald	147
903	Treptow-Köpenick	Baumschulenweg	527
904	Treptow-Köpenick	Johannisthal	438
905	Treptow-Köpenick	Niederschöneweide	388
906	Treptow-Köpenick	Altglienicke	192
907	Treptow-Köpenick	Adlershof	818
908	Treptow-Köpenick	Bohnsdorf	70
909	Treptow-Köpenick	Oberschöneweide	630
910	Treptow-Köpenick	Köpenick	1226
911	Treptow-Köpenick	Friedrichshagen	448
912	Treptow-Köpenick	Rahnsdorf	70
913	Treptow-Köpenick	Grünau	273
914	Treptow-Köpenick	Müggelheim	59
915	Treptow-Köpenick	Schmöckwitz	40
1001	Marzahn-Hellersdorf	Marzahn	474
1002	Marzahn-Hellersdorf	Biesdorf	503
1003	Marzahn-Hellersdorf	Kaulsdorf	743
1004	Marzahn-Hellersdorf	Mahlsdorf	327
1005	Marzahn-Hellersdorf	Hellersdorf	116
1101	Lichtenberg	Friedrichsfelde	741
1102	Lichtenberg	Karlshorst	1121
1103	Lichtenberg	Lichtenberg	613
1104	Lichtenberg	Falkenberg	3
1106	Lichtenberg	Malchow	1
1107	Lichtenberg	Wartenberg	28
1109	Lichtenberg	Neu-Hohenschönhausen	302
1110	Lichtenberg	Alt-Hohenschönhausen	535
1111	Lichtenberg	Fennpfuhl	28
1112	Lichtenberg	Rummelsburg	507
1201	Reinickendorf	Reinickendorf	2574
1202	Reinickendorf	Tegel	1611
1203	Reinickendorf	Konradshöhe	180

Tabelle 10: Liste der Berliner Ortsteile

Nummer	Bezirk	Ortsteil	Anahl der Beobachtungen
1204	Reinickendorf	Heiligensee	219
1205	Reinickendorf	Frohnau	476
1206	Reinickendorf	Hermsdorf	362
1207	Reinickendorf	Waidmannslust	162
1208	Reinickendorf	Lübars	122
1209	Reinickendorf	Wittenau	935
1210	Reinickendorf	Märkisches Viertel	11
1211	Reinickendorf	Borsigwalde	198

Literatur

[1] Bazyl, M. (2005). Hedonic price model for Warsaw housing market, *Department of Applied Econometrics Working Papers* No. 8-09.

[2] Bivard, R., Pebesma, J. und Gómez-Rubio, V. (2008). *Applied Spatial Data Analysis with R*, Springer, New York.

[3] Brasington, D. M. und Hite, D. (2005). Demand for environmental quality: a spatial hedonic analysis, *Regional Science and Urban Economics* 355: 57-82

[4] Brunauer, W. A., Lang, S., Wechselberger, P. und Bienert, S. (2010). Additive hedonic regression models with spatial scaling factors: An application for rents in Vienna, *The Journal of Real Estate Finance and Economics* 41: 390-411.

[5] Fahrmeir, L., Kneib, T. und Lang, S. (2009). *Regression*, Springer, Berlin.

[6] Harrison, D. und Rubinfeld, D. L. (1978). Hedonic housing prices and the demand for clean air, *Journal of Environmental Economics and Management* 5: 81-102.

[7] Investitionsbank Berlin, *IBB Wohnungsmarktbericht 2013*.

[8] Martins-Filho, C. und Bin, O. (2005). Estimation of hedonic price functions via additive nonparametric regression, *Empirical Economics* 30: 93-114.

[9] Richardson, H. W., Vipond, J. und Furbey, R. A. (1974). Determinants of urban house prices, *Urban Studies* 11: 189-199.

[10] Sheppard, S. (1998). Hedonic analysis of housing markets, *Handbook of Regional and Urban Economics* 3: 1595-1635.

[11] Theebe, M. A. J. (2004). Planes, Trains, and Automobiles: The impact of traffic noise on house prices, *The Journal of Real Estate Finance and Economics* 28: 209-234.

Erklärung zur Urheberschaft

Hiermit erkläre ich, Karolina Stańczak, dass ich die vorliegende Arbeit allein und unter Verwendung der aufgeführten Quellen und Hilfsmittel angefertigt habe. Die Prüfungsordnung ist mir bekannt. Ich habe in meinem Studienfach bisher keine Bachelorarbeit eingereicht.